J'ai l'impression que

Vandeweerd, Nathan; Keijzer, Merel

Published in:
Canadian Journal of Applied Linguistics

Publication date:
2018

J'ai l'impression que: Lexical Bundles in the Dialogues of Beginner French Textbooks

Nathan Vandeweerd
Université catholique de Louvain

Merel Keijzer
University of Groningen

Abstract

Formulaic language is notoriously difficult for second language learners of French to master (Edmonds, 2014; Forsberg, 2010). Yet, no study has examined formulaic language in French textbooks despite the fact that in many contexts, textbooks represent a significant proportion of the input that learners receive. The current study addresses this gap. Using a distributional approach (as used in Biber, Conrad, & Cortes, 2004), four-word lexical bundles were extracted from an oral corpus of French. The average number of lexical bundles in oral corpus utterances was compared to the average number of bundles in a corpus of A1-B1 level textbook dialogues. An independent samples *t* test showed that the average number of lexical bundles per 100,000 words was significantly higher in texts from the oral corpus than the textbook corpus. The average number of stance and referential lexical bundles was also revealed to be higher in the oral corpus. Implications for textbook design are discussed, such as increasing the amount of formulaic language in A2 level textbooks and incorporating more authentic language into textbooks.

Résumé

La maîtrise du langage formulaïque constitue un défi majeur pour les apprenants du français langue seconde (Edmonds, 2014 ; Forsberg, 2010). Néanmoins, aucune étude n'a examiné le langage formulaïque dans les manuels de français langue seconde, malgré le fait qu'ils représentent une proportion importante de l'input que les apprenants reçoivent. La présente étude vise à combler cette lacune. Nous avons utilisé une approche distributionnelle (comme ont utilisé Biber, Conrad et Cortes, 2004) pour extraire des groupes lexicaux de quatre mots à partir d'un corpus oral de français. Nous avons ensuite comparé le nombre moyen de groupes lexicaux dans les énoncés oraux et dans les dialogues dans les manuels A1-B1. Le nombre moyen de groupes lexicaux pour 100 000 mots était plus élevé dans le corpus oral comparé au corpus de manuels et un test *t* pour échantillons indépendants a révélé que cette différence était significative. Le corpus oral contenait aussi une plus grande proportion de groupes lexicaux de types positionnel et référentiel. Nous traitons des implications que cela engendre pour la conception de manuels, notamment l'augmentation du langage formulaïque au niveau A2 et l'incorporation du langage authentique à tous les niveaux.

J'ai l'impression que: Lexical Bundles in the Dialogues of Beginner French Textbooks

Introduction

Evidence continues to show the importance of input to second language (L2) development. Proponents of a usage-based theory of L2 development argue that learners are sensitive to the quality and quantity of input they receive (Ellis & Larsen-Freeman, 2009). In many foreign language-learning situations, the textbook is the most significant source of linguistic input for learners, so great care should be taken in the design of coursebook materials. Unfortunately, recent research has shown that most textbooks are not successful in helping students to acquire language (Tomlinson, 2016). Tomlinson (2016) argued that one reason that textbooks fall short in this regard is that they do not provide rich, authentic input reflective of the target language.

Research has also continued to show that a large proportion of speech is composed of relatively fixed expressions that have become conventionalized in a speech community. Formulaic language (FL) accounts for up to one third of spoken French, for instance (Forsberg, 2010). In addition, FL has been shown to be important in discourse organization (Erman & Warren, 2000; Granger & Paquot, 2008) and in achieving native-like accuracy (Pawley & Syder, 1983) and fluency (Kuiper, 2004). Furthermore, FL can confer processing advantages (Conklin & Schmitt, 2012; Jiang & Nekrasova, 2007; Tremblay, Derwing, Libben, & Westbury, 2011; Underwood, Schmitt, & Galpin, 2004) and learners may be able to extrapolate linguistic information from formulaic sequences of words (Ellis, 2012; Myles, Hooper, & Mitchell, 1998; Myles, Mitchell, & Hooper, 1999). Given the ubiquity and importance of FL to natural language, it follows that FL should also be present in a high proportion in learner-directed input. As textbooks constitute a major source of this input, they should thus contain a large amount of FL.

Although previous studies have examined FL in English textbooks (Biber, Conrad, & Cortes, 2004; Koprowski, 2005), thus far no study has investigated FL in French textbooks, despite the fact that for learners of French in foreign-language classroom contexts, the textbook may be the only significant source of input, much more so than for English language learning.

The purpose of the current study is to determine whether dialogues in beginner French textbooks reflect accurate and authentic use of FL and to provide insights to textbook publishers and educational decision makers, which will allow them to improve the amount and quality of FL in L2 curricula and materials.

Theoretical Background

A Usage-Based Framework

In a usage-based theory of L2 development, FL is of particular interest. Through repeated usage, certain units of language become conventionalized in a speech community. Users of a language interact using the conventionalized units and adapt to new situations by breaking those units down, thereby creating the patterns of language that we recognize as grammatical rules. Eventually, chunks such as *je m'appelle X* (my name is X) become entrenched in the mind of a speaker and in a speech community (Ellis & Larsen-Freeman,

2009). In contrast to a generative framework (cf. Chomsky, 1994), wherein grammar is composed of a series of universal constraints on language, proponents of a usage-based framework claim that grammatical patterns emerge due to the abstraction of patterns in the input. The more frequent an item is in the speech community, the more likely it is to become entrenched as a conventionalized unit. In short, FL is a key component of usage-based approaches to language learning. As language units become conventionalized, certain expressions are found to regularly occur together as "chunks" and may be processed as a whole.

Defining FL

As an indication of the diversity of phenomena covered by the label of FL, Wray and Perkins (2000) cited over 40 different terms that researchers have used, including *collocations*, *idiomatic expressions* and *prefabricated routines and patterns*. Wray and Perkins collected all of these terms under one single umbrella term, *formulaic language*. For these authors, the uniting factor was the fact that certain expressions are "stored and retrieved whole from memory at the time of use" (Wray & Perkins, 2000, p. 1). This approach assumes that there is a binary distinction between FL and non-FL. According to this definition, phrases that are stored together as a chunk are formulaic and phrases that are generated by a grammar are non-formulaic.

Using conventionality as the basis for identifying formulaic sequences surpasses the binary distinction between FL and non-FL. As a proponent of the usage-based framework, Ellis (2012) argued that the degree to which an expression is formulaic is related to the extent to which that expression is entrenched both for the speaker and in the speech community at large. Every expression is somewhere on a continuum from unconventional to conventional. In fact, processing speed of formulaic expressions was found to be associated with the frequency of the expressions in corpora (Conklin & Schmitt, 2012). According to Ellis (2012), "these findings argue against a clear distinction between linguistic forms that are stored as formulas and ones that are computed or openly constructed" (p. 25). For that reason, this paper defines formulaic sequences as units of language that are conventional to the extent that they are frequent in a speech community.

One way of measuring the frequency of FL in a speech community is by using corpus data. Biber et al. (2004) used an automatic extraction technique to identify expressions that they called *lexical bundles*. Despite the fact that the lexical bundles did not fit traditional syntactic categories, they did seem to reflect the three main discursive functions as proposed by Granger and Paquot (2008): referential expressions, which refer to real-world entities (i.e., *heavy rain*); discourse organizers (called *textual phrasems* in Granger and Paquot), which serve to organize speech (i.e., *in addition to*), and stance expressions (called *communicative phrasems* in Granger and Paquot), which are used to express opinions and attitudes (i.e., *I think that*). Data from distributional analyses of corpora such as those used by Biber et al. support usage-based theories of L2 development, showing that despite the possibly boundless nature of language, most of what people actually say is limited to a set of high-frequency sequences in a Zipfian distribution (Ellis, 2012). That is to say, the most frequent phrases account for a much larger percentage of a given corpus than the less frequent phrases.

FL and L2 Development

It has been noted that young children (initially) reproduce, as a whole, sequences of words from adult language without an understanding of the meaning of the component parts (cf. Cruttenden, 1981). Wray (2002) detailed the steps towards a more analytic use of FL in children and went on to draw parallels with L2 learning. As in first language (L1) acquisition, some beginning learners rely heavily on FL in order to communicate before they have acquired the grammatical competence to generate sentences on their own (Wray, 2002). However, FL in intermediate and advanced learners in classroom settings tends to "lag behind expectations" (Wray, 2002, p. 148) and, generally, only very advanced learners (near-native speakers) have been found to use FL accurately. According to Wray (2002), learners acquiring an L2 use fewer formulaic sequences because they have the option to rely on their L1 (through translation) when communicating. Moreover, in the classroom context, learners are mostly taught to analyze language from the early learning stages onwards, which may cause them to break down sequences into their component parts instead of remembering them as holistic expressions.

In line with what is known about L2 learners in general, for French learners, only those with a high proficiency level have been shown to be native-like in the use of FL. Forsberg and Bartning (2010) found, for instance, that the number of formulaic sequences used by Swedish learners of French at A2, B2 and C2 levels was significantly different and that the students at the C2 level used the highest number of referential FL expressions. Forsberg (2010) also found that the number of referential FL chunks could reliably distinguish between proficiency levels and that there was no difference between advanced non-native speakers and native speakers in the number of conventional sequences used in speech. Moreover, Lundell and Lindqvist (2012) found that the non-native speakers in their study were able to achieve native-like levels of referential formulaic sequences in oral production. Learners have also been found to possess receptive knowledge of which sequences are formulaic, showing that they have some sensitivity to the frequency of the expressions in the input. Edmonds (2014) elicited a set of conventional responses to common situations from native speakers of French. When exposed to the conventional expressions along with a set of length-matched control expressions, non-native speakers and native speakers alike judged the conventional expressions as more natural in the same proportion. However, reaction times to the responses of the non-native speakers were still significantly slower than those of native speakers. Even advanced non-native speakers failed to produce native-like use of FL. Bolly (2008) studied university-level English-speaking learners of French and found that they overused verb-noun collocations with the verb *donner* (to give) and underused verb-noun collocations with the verb *prendre* (to take) in their writing (c.f. Edmonds, 2014).

Longitudinal studies have shown that increased exposure to L2 input can help learners improve their use of FL (Raupach, 1984; Towell, Hawkins, & Bazergui, 1996, which is consistent with a usage-based model of L2 development, so incorporating FL into pedagogic materials may be one way to accomplish this.

Incorporating FL in Pedagogic Materials

In their discussion of pedagogic interventions to teach FL, Boers and Lindstromberg (2012) pointed out that acquisition of formulaic sequences is "strongly contingent upon the frequency of occurrence of the items in the input" (p. 99). Moreover, Tomlinson (2016) argued that textbooks should provide students with "rich, re-cycled, meaningful and comprehensible input of language in use" ("Prerequisites for Language Acquisition", para. 2). To that end, several researchers have investigated the extent of FL in textbooks and graded readers directed towards English language learners. Tsai (2015) studied the collocational profile of three English textbooks commonly used in Taiwan, finding that the density of collocations was significantly different in each textbook. In addition, the collocational density of all three textbooks was small. Webb, Newton, and Chang (2013) suggested repeating collocations up to 15 times for optimal learning but Tsai found that 90% of the collocations in the textbooks only occurred once. As he pointed out though, it would be impossible to expect textbooks to incorporate all possible collocations but it is important to strike a balance between usefulness and frequency in natural language. Tsai used a top-down approach because the collocations generated by bottom-up methods, which compile collocation lists from large corpora such as the British National Corpus (BNC), "may be too uncommon in real language use" (Tsai, 2015, p. 725). For example, though the collocation "to use a/the textbook" is frequent in a school setting, its frequency in the BNC is not statistically significant. However, problems of pedagogic relevance may be solved by using corpora that are characteristic of the target language.

Biber et al. (2004) showed that a bottom-up method can indeed be useful if the corpus from which FL is extracted represents the target language. They compared lexical bundles in a corpus of academic speaking and writing to university textbooks and classroom language. Looking at the number of lexical bundles in each register, Biber et al. found that not only did the spoken registers have more lexical bundles than written registers but the proportion of discourse organizers and stance organizers was much higher in both conversation and classroom teaching than in textbooks. As Biber et al. pointed out, "it is surprising that textbook authors do not incorporate more lexical bundles in their writing, given the heavy reliance on bundles in classroom teaching" (p. 383). Although the textbooks chosen for Biber et al.'s study were subject-specific textbooks rather than language teaching textbooks, the difference in both the number and type of lexical bundles is relevant to the question of whether the input students receive from text-based materials will allow them to acquire native-like fluency and native-like selection of formulaic sequences (Pawley & Syder, 1983). According to the results of Biber et al., students exposed primarily to textbooks would not receive the input required to develop proficient use of lexical bundles in speech.

Koprowski (2005) studied FL in intermediate level English as a foreign language textbooks, finding that more than 14% of the lexical phrases in the textbooks were not found in the COBUILD corpus. Furthermore, 23% of the phrases had an extremely low frequency in the corpus and a limited range across subcorpora. There was also less than 1% agreement between textbooks. According to Koprowski (2005), "designers did not structure their course around a body of useful lexical phrases, but rather, started in most cases with a theme, topic, or structure and then considered items related to these basic concepts" (p. 330). This was also found by Gouverneur (2008), who studied the collocations including the words *make* and *do* in a corpus of 10 advanced level and seven intermediate level

English Language Teaching textbooks. Gouverneur found little agreement between textbooks as to which collocations were included. Taken together, these findings suggest that the FL found in textbooks is rather impoverished. In foreign language learning contexts, the textbook may be the single most important source of input for students. Given such a situation, it is important for textbook publishers to ensure that their coursebooks contain authentic FL.

Although there is work relating to FL used in English Language Teaching textbooks, thus far, no study has addressed the extent of FL in French textbooks. The current study addresses this gap by examining FL in dialogues from beginner French textbooks. The goal of the study is to answer the following research question: To what extent is the FL in the dialogues of beginner French textbooks representative of authentic language use? In order to answer this question, the following subquestions are also addressed:

1. What are the most frequent lexical bundles in a corpus of spoken French?
2. What discursive roles do the above lexical bundles play?
3. Is there a significant difference between utterances in the oral corpus and dialogues in the textbook corpus with respect to the average number of lexical bundles?

Methodology

The Oral Corpus

Both an oral corpus and a textbook corpus were compiled for the purpose of this study. In order to ensure that the two corpora were comparable, the following criteria were set for the selection of the oral corpus to be used. First, the corpus needed to contain casual speech by native French speakers on topics that learners of French at A1-B1 level are able to discuss (the same level as the textbooks in the textbook corpus). According to the Common European Framework of Reference (CEFR), a learner at this level can ask and answer questions about personal habits and routines, what they do at work and in their free time, and provide personal information (Council of Europe, 2001). Because the textbooks used in the current study were aimed at teenage and adult learners, the corpus also needed to contain speech by speakers of a similar age range. Finally, the data from the oral corpus also needed to be freely downloadable so that they could be analyzed using CLAN software (MacWhinney, 2012). Although there are French oral corpora that meet the content criteria (e.g., Branca-Rosoff, Fleury, Lefeuvre, & Pires, 2012; CNRS, 2016; Laboratoire ICAR, 2014), either they do not allow transcripts to be downloaded or the size of any given corpus was too large to be workable for the current project.

The first corpus of spoken French that was selected for use in the current study was the French Interlanguage (InterFra) corpus, compiled at Stockholm University (https://spraakbanken.gu.se/swe/resurs/interfra). The corpus contains native speakers and non-native speakers performing various tasks in French but for the current study, only the interview data from native speakers were used. The interviews were conducted by a native speaker and covered topics related to family, employment, interests, hobbies, school, and the cultural differences between Sweden and France. The topics are comparable to the topics that a learner studying French at the A1-B1 level can discuss (Council of Europe,

2001). The average age of the native speakers in the InterFra corpus ranges from 21-52 and the total number of words in the InterFra corpus is 111,372. However, as the textbooks used in the current study are targeted towards both teenage and adult learners and the youngest speaker in the InterFra corpus is 21, a second corpus of teenage speech was also used.

The System-Aided Compilation and Open Distribution of European Youth Language corpus (SACODEYL) is an open-source corpus of spoken youth language in Europe (Chambers, 2009). The interviews in the SACODEYL French corpus are similar in topic to the InterFra interviews and CEFR guidelines for A1-B1 level students (Council of Europe, 2001). Topics include family, hobbies, travel, and school life. The corpus contains 22 one-on-one interviews and two group interviews. Age information was not available for every speaker but based on group interviews with students in the same class, age was estimated for those who did not explicitly state their age. The ages of the students in the corpus range from approximately 11 to 17 years (M = 14, SD = 2.1). The SACODEYL corpus is divided into two age groups according to the school levels in the French education system: students in *college* (junior high school), aged 11-14 years (n = 12) and students in *lycée* (high school), aged 15-17 years (n = 12). Both groups were native French speakers; however, the *collège* students were residents of Guadeloupe, a French overseas territory whereas the *lycée* students lived in mainland France. With the addition of the SACODEYL corpus, the total number of words in the combined oral corpus was 154,910.

Both corpora were prepared for analysis by removing the speaker codes before each line of speech and removing all line breaks. All non-orthographic characters, ellipses, and question marks were likewise removed. Spelling conventions were standardized across both corpora (e.g., *tee shirt* for *t-shirt*, *pourcent* for *pour cent*). Errors in the InterFra corpus from incompatible Unicode coding were corrected and all morphological tagging was removed. In addition, the text of the interviewer was absent in some InterFra interviews so it was re-transcribed from the audio files available on the InterFra website. All integers were re-coded as "NUMi" and all ordinal numbers were re-coded as "NUMo." All conjunctions were separated (i.e., *j' --> je*). To facilitate the process, the replacement was carried out on all files simultaneously using TextWrangler Software (Bare Bones Software Inc., 2016). The conjunction *l'* can represent the feminine or masculine definite article but all instances of *l'* were re-coded as *le* to facilitate automated re-coding.[1] As demonstrated by Forsberg (2010), many formulaic sequences occur in several inflections: *ai peur de* ([I] am scared of), *as peur de* ([you] are scared of), *a peur de* ([s/he] is scared of) [p. 46]. Thus, all inflections of the two auxiliary verbs *avoir* (to have) and *être* (to be) were re-coded as the infinitive forms AVOIR and ETRE. All reduplications of four, three, two, and one words were deleted in order to avoid speaker fluency errors from influencing the overall count of lexical bundles (e.g., *je pense que je pense que --> je pense que*).

The Textbook Dialogue Corpus

For the textbook dialogue corpus, the same textbooks that were used in François (2011, 2014) were used to construct the corpus in the current study for the following reasons. First, one of the inclusion criteria for François (2011, 2014) was that the textbooks had to be compatible with the CEFR and thus were published (with the exception of the *Panorama* series of textbooks) since 2001. The CEFR designation of the publisher was taken as the main proficiency level criterion. In some cases, however, a textbook contained lessons intended for more than one CEFR level, in which case the chapters were assigned

separate CEFR levels. This criterion is compatible with the current study to the extent that the interview topics in the oral corpus consist of tasks at the A1-B1 level (Council of Europe, 2001). Second, the textbooks are targeted towards teenagers or young adults learning French. Finally, the textbooks in François (2011, 2014) are designed for general French as a foreign language and are not targeted towards learners with a specific L1 background. There are no French for specific purposes textbooks in the corpus. The three main publishers whose textbooks were used in François (2011, 2014) are European publishers and it is assumed that the primary audience of the textbooks are students studying French as a foreign language in Europe. François (2011, 2014) found 24 textbooks that met the above criteria and from which texts of nine genres[2] were extracted. Because the research aim of François (2011, 2014) was readability, oral exercises were not included. However, the corpus does include dialogues, which makes it useful for the current study. Since the textbooks collected for inclusion in the corpus created by François (2011, 2014) were all published prior to 2008, the corpus in this study was expanded by including dialogues from textbooks published since 2008. The original corpus contains textbooks from four publishers: CLE International, Hachette, Didier, and Difusión. To ensure that the new textbooks would be comparable to the textbooks assembled by François (2011, 2014), one new textbook series each from CLE International, Hachette, and Didier were chosen for inclusion in the new corpus.[3] The same criteria were used in the selection of the post-2008 textbooks: dialogues from general-purpose French as a foreign language textbooks targeted towards no particular L1 were included. As the original corpus contains only five texts beyond the B1 level, only textbooks representing levels A1-B1 were included in the final corpus. Following François (2011, 2014), only dialogues containing at least two speakers were included. No materials from workbooks or teacher handbooks were included. The transcriptions of the dialogues were scanned and the text was extracted using Google Drive-integrated OCR software (Google, 2016). All dialogue prefaces and speaker tags were removed and the re-coding operations described above were carried out using TextWrangler (Bare Bones Software Inc., 2016). The resulting combined textbook corpus is almost equally divided between pre-2008 (34,576 words) and post-2008 textbooks (44,890 words) and between proficiency levels (A1: 26,555 words; A2: 25,647 words; B2: 27,264 words). The one exception to this is that the Echo Series published by CLE International spreads the B1 level over two textbooks, which means that the size of the post-2008 B1 subcorpus is slightly larger than the pre-2008 B1 subcorpus.

Extracting Lexical Bundles

Following Suethanapornkul (2009), lexical bundles were extracted using CLAN software (MacWhinney, 2012). With the COCCUR command, CLAN extracts the most frequent phrases of *n* length from a list of text files without crossing sentence boundaries. In order to limit the scope of the study, and to be consistent with Biber et al. (2004), a length of four words was chosen. It should be noted, however, that Biber et al. pointed out that setting the cut-off at four words is arbitrary and that sometimes lexical bundles can combine together to form longer bundles. The command was executed on the %TXT tier, which represents the surface representation without grammatical coding, as was done in Suthanapornkul (2009). However, as previously mentioned, the grammatical inflections of the two most frequent French verbs were collapsed because Forsberg (2010) had previously shown that many formulaic sequences in French can occur in inflected forms. The exact

command used was thus the following: cooccur +t*txt +n4 +d1 +u +o *.cha. The command tells CLAN to search for re-occurring four-word sequences in the text files. It also tells CLAN to search all the text files together and to export the data to an Excel file in order of descending frequency. In order to be included in the final list of lexical bundles, the candidate sequences needed to meet the following criteria. First, the sequence had to occur at least four times in each corpus and at least once in five different samples. As Biber et al. (2004) pointed out, the establishment of a threshold is rather arbitrary. The current study thus uses the same threshold established by Biber et al. of 40 times per 1,000,000 words. However, given the smaller size of the corpus in this study, the proportional value of four times per 100,000 words was chosen. The criterion of occurrence in at least five different texts was set as in Biber et al. to guard against idiosyncratic use by a single speaker. Finally, each lexical bundle needed to occur at least once in both the InterFra and SACODEYL corpus. To further limit the scope of the current study, a threshold of the first 100 candidates matching this description was set. However, because the frequency of occurrence per 100,000 words of some lexical bundles was the same, the final list contained 103 lexical bundles.

Once the list of lexical bundles was compiled, the bundles were categorized by discursive function according to the schema developed in Biber et al. (2004). Similar schemas for categorizing FL have been developed by Granger and Paquot (2008) and Forsberg (2010) but these schemas categorized formulaic sequences by syntactic function and as shown by Biber et al., lexical bundles often do not respect traditional syntactic boundaries. For this reason, the Biber et al. schema was used. This schema organized lexical bundles into three main types: (a) stance expressions, which "express attitudes or assessments of certainty"; (b) discourse organizers, which "reflect relationships between prior and coming discourse," and (c) referential expressions, which "make direct reference to physical or abstract entities, or to the textual context itself" (Biber et al., 2004, p. 384). In addition to these three categories, Biber et al. also categorized a small minority of lexical bundles as "special conversational" bundles. These included politeness markers, inquiry, and reporting functions. In the current study, lexical bundles were classified into the same four categories. Biber et al. also mentioned that some lexical bundles may serve more than one function. Thus, in order to classify the bundles, three randomly generated samples were read containing each bundle and a primary function was assigned to the lexical bundle according to the function that the bundle appeared to serve in two out of the three cases. Because each text in the corpus was of a different length, the total number of lexical bundles and the number of lexical bundles in each discursive category in each text was standardized to the number of lexical bundles per 100,000 words.

Analysis

Following Gries (2006), before the oral and the textbook corpus were compared, the data were first normalized in order to examine the homogeneity of each corpus. As the number of texts in each subcorpus was different, a Kruskall-Wallis H test was used to compare the mean number of lexical bundles per 100,000 words between subcorpora. Once the homogeneity of the subcorpora was assured, the oral and textbook corpora were compared with respect to the average number of lexical bundles per 100,000 words using an independent samples *t* test. The average number of each type of lexical bundle (stance, discourse, referential, communicative) was also compared between the corpora using

independent samples *t* tests. Significant differences between the corpora with respect to the number of specific lexical bundle types were further investigated in the textbook corpus using a three-way step-wise Anova with three factors: date of publication (pre/post 2008); level (A1, A2, B1); and publisher (Hachette and CLE International).[4] This was done in order to determine: (a) whether there was a significant difference between textbooks published before and after 2008; and (b) whether textbooks from each publisher varied significantly in the average number of stance, referential, discourse, and communicative bundles. As the number of samples in each group was unequal, a Hochberg GT2 posthoc test was used. An alpha decision level of 0.5 was chosen as is conventional.

Results

From the combined oral corpus, a total of 103,292 four-word lexical bundles were initially extracted, accounting for 77.72% of the entire oral corpus. The 103 most frequent bundles matching the criteria outlined above were used for further analysis, comprising 1.77% of the entire corpus. A Kruskal-Wallis H test showed that there was no statistically significant difference in the average number of lexical bundles in each subcorpus [$X^2(5) = 0.48$, p = 1.000].

As in Biber et al. (2004), the lexical bundles were categorized according to their function. Of the 103 lexical bundles selected for investigation, there were 34 (33%) referential bundles (e.g., *ce ETRE pas vraiment*, it isn't really); 42 (41%) discourse bundles (e.g., *ce ETRE à dire*, that is to say); 24 (23%) stance bundles (*je crois que ce,* I believe that); and three (3%) special communicative bundles (*je AVOIR NUMi ans,* I am X years old; *je me ETRE dit*, I said to myself; *oui ce ETRE ça*, yes that's right). The textbook corpus was also divided into three subcorpora based on publisher. A Kruskal-Wallis H test revealed that there was no statistically significant difference in the average number of lexical bundles in each subcorpus [$X^2(2) = 5.743$, $p = .057$] and so the textbook corpus could be compared as a whole to the oral corpus. As shown in Table 1, an independent samples *t* test showed that the average number of lexical bundles was higher in texts from the oral corpus ($M = 1733.65$, $SD = 642.12$) when compared to the textbook corpus ($M = 1274.57$, $SD = 1942.53$). This difference was significant [$t(214.2) = 3.9$, $p < .001$]. On average, the oral corpus utterances also contained more of each type of lexical bundle (stance, discourse, referential, communicative) than the textbook dialogues but this was only significant for the referential bundles [$t(269.31) = 2.12$, $p = .035$] and stance bundles [$t(108.75) = 5.37$, $p < .001$].

Table 1
Mean Number of Lexical Bundles per 100,000 Words in Oral and Textbook Corpora (Standard Deviation in Parentheses)

Corpus	Overall	Stance	Referential	Discourse	Communicative
Oral	**1733.65 (642.12)	**392.66 (297.09)	*525.26 (269.46)	779.11 (332.88)	35.91 (46.50)
Textbook	**1274.57 (1942.53)	**151.91 (542.83)	*412.35 (937.76)	680.96 (1226.48)	29.35 (168.92)

*$p < .05$. **$p < .001$.

As the difference between the mean number stance bundles in the oral and textbook corpora was significant, the textbook corpus was examined as to the role of date of publication, level and publisher. In the case of stance bundles, there was a significant effect of date of publication on the number of stance bundles [$F(1, 493) = 4.07$, $p = .044$]. The average number of stance bundles words was higher in textbooks published after 2008 ($M = 230.60$, $SD = 704.95$) than in textbooks published before 2008 ($M = 72.65$, $SD = 287.41$). An independent samples t test confirmed that this difference was significant [$t(342.85) = -3.3$, $p < .001$)]. The effect of level was not significant [$F(2, 493) = 2.13$, $p = .120$] nor was the effect of publisher [$F(1,493) = 0.652$, $p = .420$]. There were no other significant isolated or combined effects regarding the average number of stance bundles in textbook dialogues.

The influence of date of publication, level, and publisher on the average number of referential bundles was also examined in the textbook corpus. There was a significant effect of level [$F(2, 493) = 3.53$, $p = .030$]. Hochberg GT2 posthoc test revealed that the difference between A1 level ($M = 239.65$, $SD = 534.78$) and A2 level ($M = 403.62$, $SD = 911.13$) textbooks was not significant ($p = .253$). However, the difference between A1 and B1 level ($M = 802.80$, $SD = 1423.38$) was significant ($p < .001$) as was the difference between A2 and B1 level textbooks ($p = .001$). There was also a significant combined effect of date of publication and level for the number of referential bundles [$F(2,493) = 4.90$, $p = .008$]. Independent samples t tests revealed that the difference between the pre- and post-2008 textbooks was not significant in A1 level textbooks [$t(153.56) = -0.604$, $p = .547$] or the A2 level textbooks [$t(144.66) = 1.577$, $p = .117$] but pre-2008 textbooks had fewer referential bundles ($M = 110.35$, $SD = 258.20$) at the B1 level than the post-2008 textbooks ($M = 963.67$, $SD = 1532.28$) and this difference was significant [$t(115.72) = -5.23$, $p < .001$] (cf. Figure 1). Finally, there was a significant combined effect of level and publisher on the number of referential bundles [$F(2,493) = 3.44$, $p = .033$]. Independent samples t tests revealed that the difference in the number of referential bundles between the publishers Hachette and CLE International were not significant in A1 level textbooks [$t(121.47) = -0.95$, $p = .925$] or B1 level textbooks [$t(54.72) = -0.749$, $p = .457$] but the difference was significant in A2 level textbooks [$t(39.74) = 2.12$, $p = .040$] (Figure 2). In other words, on average, dialogues in A2 level textbooks published by Hachette had more referential bundles ($M = 800.47$, $SD = 1529.88$) than those in A2 level textbooks published by CLE International ($M = 265.27$, $SD = 495.66$) [Figure 2]. No other significant isolated or combined effects were found.

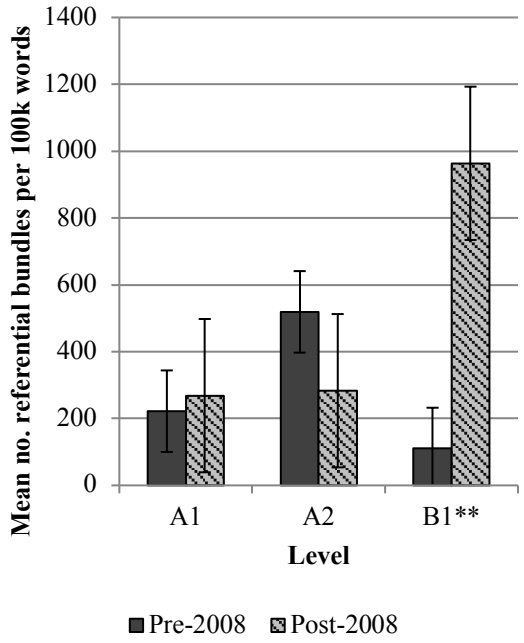

Figure 1. Mean number of referential bundles by level and date of publication (error bars +/- 2 SE, $^{**}p < .001$).

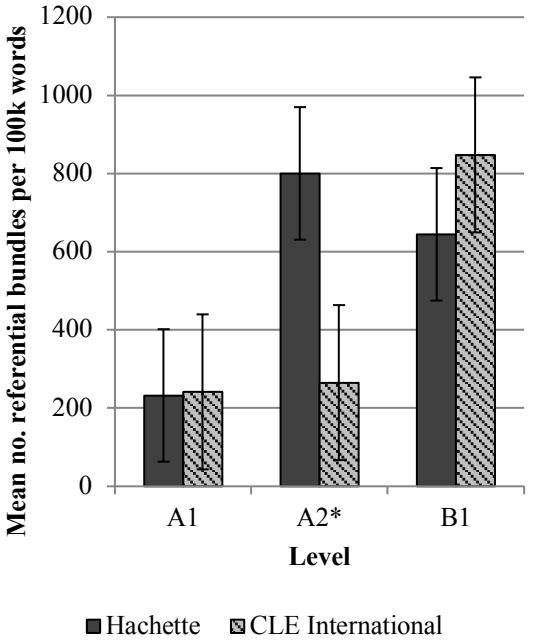

Figure 2. Mean number of referential bundles by level and publisher (error bars +/- 2 SE, $^{*}p < .05$).

To summarize, 103 high-frequency lexical bundles were extracted from the oral corpus and categorized according to four discursive functions: referential, discourse, stance, and communicative. The most frequently occurring bundles in both corpora are provided in Table 2. On average, utterances in the oral corpus had a higher number of lexical bundles

than dialogues from the textbook corpus. The average number of referential, discourse, stance, and communicative bundles was also lower in the textbook corpus but this difference was only significant for stance bundles and referential bundles. Upon closer inspection within the textbook corpus, the number of stance bundles was found to be significantly greater in the post-2008 textbooks, showing that the more recent textbooks have a higher number of stance bundles overall (but still lower than the levels found in the oral corpus). Regarding referential bundles, there was an effect of textbook level on the average number of referential bundles but this effect interacted with the date of publication and the textbook publisher. Whereas the pre-2008 textbooks show a downward curve in the number of referential bundles at the three levels, with the highest number in the A2 level textbooks and the lowest number in the B1 level textbooks, the number of referential bundles in post-2008 textbooks increases somewhat gradually at each level. This pattern is also mirrored by the textbooks published by CLE International, which in contrast to Hachette, gradually increases the number of referential bundles at every level. Taken together, these findings indicate that there is a lower number of lexical bundles overall in the textbook dialogues when compared to real interactions between native speakers and that there is a difference between textbooks published by different publishers with regard to the number of different types of lexical bundles at each level.

Table 2
Most Frequently Occurring Lexical Bundles (>40 per 100,000 words), Listed in Decreasing Frequency in the Oral Corpus

Lexical Bundle	Type	Oral Corpus (Freq./100k words)	Textbook Corpus (Freq./100k words)
ETRE ce que tu	discourse organizer	174.29	0
que ETRE ce que	discourse organizer	173.65	8.79
ce ETRE vrai que	stance	117.48	22.61
que il y AVOIR	referential	83.27	2.51
ce ETRE à dire	discourse organizer	78.75	85.42
ETRE ce que vous	discourse organizer	61.33	27.64
ce ETRE un peu	referential	59.39	40.20
ce que tu AVOIR	discourse organizer	49.71	30.41
parce que ce ETRE	discourse organizer	47.12	188.44
ETRE a dire que	discourse organizer	43.31	26.38
y AVOIR beaucoup de	referential	41.31	13.82
il y AVOIR des	referential	40.67	42.71
et que ETRE ce	discourse organizer	40.02	8.79
parce que je AVOIR	discourse organizer	28.40	45.22
je AVOIR le impression	stance	16.78	43.97
ETRE un petit peu	referential	16.78	42.71

Discussion and Conclusion

The aim of this study was to determine the extent to which FL in the dialogues of beginner French textbooks is representative of authentic language use. This was further broken-down into three subquestions:

1. What are the most frequent lexical bundles in a corpus of spoken French?

Using a distributional method, 103,292 four-word lexical bundles were extracted from the oral corpus. Of course, not all of the extracted items are formulaic in the sense of being conventional in the speech community. In fact, a large number of these combinations are likely due to random patterns of co-occurrence. To ensure that the lexical bundles on the list were indeed conventional, the list only includes bundles that occurred at least four times in both the InterFra corpus and the SACODEYL corpus and that appeared in at least five different texts. The conventionality of the lexical bundles on this list (Table 2) makes intuitive sense given that they are used for communicative actions that occur relatively frequently. For example, the list contains simple interrogatives (*ETRE-ce que tu/vous*, do youT-form/youV-form); subordinating conjunctions (*parce que ce ETRE*, because it's) and locatives (*il y AVOIR*, there is); as well as phrases for expressing an opinion (*ce ETRE vrai que*, it is true that). Not only does the list make intuitive sense, but several of the lexical bundles on the list were also identified by Forsberg (2008, 2010) in a phraseological analysis of a spoken corpus of French, which means that the bundles identified are indeed likely to be highly conventional in spoken French.

2. What discursive roles do the above lexical bundles play?

The 103 most frequent lexical bundles in the oral corpus were categorized according to four discursive functions: referential (33%), discourse organizer (41%), stance (21%), and special communicative (3%) using the schema developed by Biber et al. (2004).

3. Is there a significant difference between utterances in the oral corpus and dialogues in the textbook corpus with respect to the average number of lexical bundles?

A comparison of the average number of lexical bundles in the oral corpus utterances and the textbook dialogues revealed that the oral corpus had, on average, more of the high frequency lexical bundles, indicating that the textbook dialogues were not representative of naturalistic speech in this regard. This mirrors previous studies that have shown that textbooks contain low levels of FL overall (Koprowski, 2005; Tsai, 2015; Webb et al., 2013), especially when compared to the amount of FL in speech (Biber et al., 2004).

The difference between the textbook corpus and the oral corpus was found to be mostly due to lower numbers of two types of lexical bundles: stance and referential. A closer look at these types of bundles in the textbook corpus seems to indicate that textbook publishers need to include more of these bundles, especially at beginner levels. Learners at this level may use FL to bootstrap their developing grammatical competence (Myles et al., 1998, 1999; Wray, 2000). As the data reveal, the number of stance bundles was not significantly different between A1, A2, and B1 level textbooks. This could mean that textbook publishers have made an effort to include stance bundles at all levels. But given that the average number of stance bundles in textbook dialogues bundles is quite low (M = 151.91, SD = 542.83) in comparison to the oral corpus (M = 392.66, SD = 297.09), this is unlikely to be the case, especially considering that recent research has shown that pragmatic language in textbooks is typically underrepresented and inauthentic (Ishihara & Paller, 2016). This suggests that either textbook publishers are unaware of the need to include

stance bundles in beginner level textbooks or that their attempt to do so has been unsuccessful.

In contrast, the number of referential bundles was significantly different in the three textbook levels but this was mostly due to the fact that dialogues from B1 level textbooks published after 2008 had a significantly higher number of this type of bundle. Textbooks published by both Hachette and CLE International include many referential bundles at the B1 level. Even so, there are still qualitative differences between the type of referential bundles in the oral corpus and the textbook corpus. For example, aspectual referential bundles (e.g., *ce ETRE pas vraiment*, it's not really) are still more common in the oral corpus than the B1 textbooks. The B1 textbooks also contain fewer imprecision bundles (e.g., *des choses comme ça,* things like that) and fewer bundles that use only one negation marker (e.g., *ce ETRE pas du,* it's not; *il y AVOIR pas,* there's not) which is common in colloquial French (Coveney, 2002), meaning that the oral corpus may contain more informal speech than the textbook dialogue corpus. The main difference between Hachette and CLE international was in the A2 level textbooks. Dialogues from textbooks published by Hachette have a significantly higher number of referential bundles at the A2 level than those from textbooks published by CLE International, which means that Hachette was more in line with research showing the importance of providing FL to beginner-level learners.

Implications for Textbook Design

One of the best ways to ensure that textbooks include FL seems to be to use authentic language samples. The textbook dialogues with the highest number of lexical bundles are often segments taken from television or radio. In fact, the dialogue with the highest number of referential bundles (6/221 words) was an authentic speech sample from the radio program *France Inter* (Figure 3, left). This dialogue is from the B1 level of the *Echo* series published by CLE International. Textbooks from this publisher had a high number of referential bundles at the B1 level. When compared to the sample from the oral corpus (Figure 3, right), it is indeed clear that both dialogues have a high number of referential bundles (especially *il y a*, there is). This makes sense given that an interview is a strenuous task that requires online planning and there is substantial research to show that FL has a beneficial effect on language processing (Conklin & Schmitt, 2012; Jiang & Nekrasova, 2007; Tremblay et al., 2011; Underwood et al., 2004) and fluency (Kuiper, 2004). What is also obvious in both examples is the presence of false starts and reformulations. The speaker in the textbook dialogue in Figure 3 (left) starts the sentence describing something: "*il y a une espèce de . . .*" (there is a type of) and then needs to think of the words, "*comment dirais-je*" (how would I say) before continuing with a new thought, "*On est contents de le faire chaque année*" (We are happy do it every year). Such a false start is also present in the oral corpus example in Figure 3 (right): "*À côté de ça je je y y je crois qu'**il y a** vraiment les les deux versants*" (Besides that I I there there I believe that there are really two sides). According to Clark (1996), repetitions and false starts of this kind are typical in natural speech as speakers need to strike a balance between fluency and accuracy. What is especially interesting in Figure 3 (right) is that the speaker uses two lexical bundles that were identified in the oral corpus: *je crois que* (I think that: stance bundle) and *il y a* (there is: referential bundle). These types of "editing expressions" are indications that the utterance is being reformulated (Clark, 1996). In the same way as the auctioneers in Kuiper (2004), these speakers seem to be relying on FL to increase fluency

in an interview situation. Speakers of an L2 also benefit from the processing advantages of FL (Jiang & Nekrasova, 2007; Tremblay et al., 2011; Underwood et al., 2004) so providing more FL in textbooks could help them increase speech fluency as well.

Patrick Boyer: Josiane nous appelle de Toulouse. Bonsoir, Josiane.
Josiane: Bonsoir, merci. Je voulais vous rappeler que le chemin de Saint-Jacques-de-Compostelle que nous faisons, nous avec mon mari, très partiellement chaque année, de 200 à 250 km, est une... comment dirais-je... ça veut dire, le champ des étoiles. Donc on a vraiment l'impression d'être un intermédiaire entre la Terre sur laquelle on prend son assise et le Ciel... Parce que le chemin de Saint-Jacques-de-Compostelle est une récupération chrétienne qui était comme l'a dit le monsieur tout à l'heure un chemin néolithique et qui permettait d'aller au bout du monde. Et **c'est un petit peu** ça, moi je ne **suis pas du tout** d'accord pour le dolorisme, mais certes, nous, on est croyants mais je pense **que c'est pas** tellement ça qui est important... C'est d'arriver au bout... **Il y a une** espèce de... Nous, de le faire de petits bouts en petits bouts... **Il y a une** espèce de... comment dirais-je de... On est contents de le faire chaque année, d'en faire un petit bout chaque année, d'arriver au bout, de rencontrer des gens qui le font pour des raisons très, très, très différentes. Et surtout dire **qu'il y avait** aussi beaucoup de symbolisme dans ce chemin et que l'on rencontre à chaque église et qu'il existe de manière très différente.

Ben je dirai que ça a un côté très le côté prise en charge organisation et tout ça. c'est vrai que c'est c'est agréable quand on arrive dans un pays et qu'on connaît absolument rien. À côté de ça je je y y je crois **qu'il y a** vraiment les les deux versants les deux côtés parce qu'en même temps **il y a des** fois où ça me où je me sens un petit peu un petit peu oppressée par ce côté prise en charge et le petit côté intégration à tout prix qui moi me qui moi me gêne me dérange. Je suis jsu je suis vraiment pas habituée à ça quoi. mais par contre je je sais **qu'il y a des** des Suédois qui qui vont partir étudier en France à Caen. les pauvres je les plains quoi. vraiment j'ai peur pour eux parce qu'il n'y a absulument aucune organisation. c'est fabuleux quoi. donc je je sais pas enfin bon je crois **qu'il y a** vraiment les les deux versants quoi.

Figure 3. Left: dialogue excerpt from the textbook corpus (Giradet & Pécheur, 2010b, p. 143). Right: excerpt from the oral corpus (InE.001), referential bundles in bold type.

In contrast, textbook dialogues with low levels of FL are rarely samples of authentic language. Instead of representing naturalistic language, these dialogues seem to be written in order to illustrate specific language features. Figure 4 is an example of a textbook dialogue, also taken from a textbook published by CLE International but at the A2 level.

M. Andriavolo: Bonjour madame Mirmont. Bonjour monsieur Issifi. Je suis très heureux de faire votre connaissance.
Laura: Nous aussi. C'est très gentil d'être venu à l'aéroport.
M. Andriavolo: Vous avez fait bon voyage?
Laura: Excellent.
M. Andriavolo: Alors bienvenue à Nosy Be. Qu'on appelle l'île aux parfums.
Tarek: C'est tout un programme. C'est ici que vous avez vos plantations?
M. Andriavolo: Mes petites plantations. Justement, je suis surpris.
Laura: De quoi monsieur Andriavolo?
M. Andriavolo: Pour acheter vos fleurs c'est moi que vous choisissez moi un petit producteur et pas la Sodexport. C'est étonnant.
Laura: On cherche un partenaire commercial c'est vrai. Mais on veut aussi quelqu'un qui participe à la création de nos parfums.

Figure 4. Example of a textbook dialogue with no referential bundles (Giradet & Pécheur, 2010a, p. 38).

This example was found to have a significantly lower number of referential bundles. As the opposing page of the textbook indicates, the goal of the exercise was to learn welcoming language. This dialogue also continues the story of the main character's perfume company. So rather than being built around natural language, this textbook series is built around a story and language elements are incorporated into dialogues ad-hoc as needed. Consequently, the language in the dialogues is not representative of natural speech. Taken together, these findings lend support to Tomlinson's (2016) first principle in second language acquisition material design: that students should be exposed to "rich, re-cycled, meaningful and comprehensible input of language in use" ("Prerequisites for Language Acquisition", para. 2). Of course, textbook authors have many criteria to keep in mind when deciding what language samples to include and so it is understandably difficult to find authentic samples of language at the right developmental level that include examples of language elements on which they want students to focus. However, given that the two textbooks samples in Figures 3 and 4 are both from CLE International, there is clearly some awareness on behalf of this publisher of the importance of including authentic input, at least at the B1 level, but this should also be extended to the A2 level textbooks because even beginner learners can benefit from FL (Wray 2000), provided it is properly scaffolded. At this developmental level, learners are also beginning to express their opinion, so having access to more pragmatic language is very useful. Providing input that contains highly conventional stance bundles, such as *c'est vrai que* (it's true that) and *j'ai l'impression que* (I have the impression that) can aid beginner learners who may use these phrases to bootstrap their developing grammatical competence. Using these phrases can also help more advanced learners overcome the naturalness problem identified by Pawley and Syder (1983). In fact, Ishihara and Paller (2016) suggested that using corpus-based conversations in textbooks may be a more effective way of teaching pragmatic language as it not only exposes students to authentic pragmatic language, but also provides the appropriate context. Though the main aim of the current project was not to compare textbooks for the proportion of authentic speech samples and dialogues written for educational purposes, these initial findings seem to suggest that the amount of FL in authentic materials is higher than in purpose-written dialogues as shown by the examples above. However, the current project was exploratory in this respect and further research should reveal whether authentic

materials and purpose-written textbook dialogues are significantly different with respect to the number of lexical bundles.

Limitations

One drawback of this study is that although the content of the interviews in the oral corpus is representative of most topics that learners should be able to discuss at CEFR levels A1-B1, some topics such as shopping or problems encountered while travelling are not extensively discussed in the interviews. The size of the corpus is also substantially smaller than the corpus used in Biber et al. (2004). Forsberg (2008, 2010) advised against using the distributional method for such a small corpus. However, given the similarity of topic coverage between the oral corpus and the CEFR guidelines, the distributional method was nonetheless employed in this case. The choice of this method is supported by the many lexical bundles that were identified both in the current study and Forsberg (2008). Nonetheless, the size and scope of the oral corpus used in the current study is indeed a limiting factor.

Likewise, although every attempt was made to ensure that the textbook corpus was well balanced with respect to level, year of publication, and publisher, it was difficult to assure that these categories were completely balanced. For example, the pre-2008 corpus did not contain textbooks published by Didier at the A2 or B1 levels and thus to ensure balance, no new textbooks at those levels were included in the post-2008 corpus. In addition, textbooks from each publisher were selected based on the general criteria of target audience and level. Due to time and access limitations, it was impossible to include every textbook from the three publishers. Style differences between textbook series published by the same publisher were also not taken into account. Furthermore, only texts that were part of the main student textbook were included.

Based on previous findings by Forsberg (2008), the decision was made to lemmatize the verbs *avoir* and *être* in the corpus data. Only these two high-frequency verbs were lemmatized so the frequency of lexical bundles containing these two verbs is naturally higher than bundles whose inflections were not collapsed. The time required to replace all instances with the infinitive and check for errors was extensive but it may have been beneficial to collapse the inflections of other high frequency verbs as well.

Finally, the task of organizing lexical bundles into discursive roles is highly problematic as discussed earlier. Many bundles can perform multiple functions and classification of the bundles is thus highly subjective, so having more raters would have also been beneficial.

The current study was primarily exploratory in nature, combining several threads of previous research on FL. It was the first study to use a distributional method to extract lexical bundles from a corpus of spoken French. Because the corpus used in the current study was small in comparison to other studies that have used a distributional approach, it would be fruitful to carry out a distributional analysis of other French corpora to confirm the conventionality of the lexical bundles extracted in the current study.

To sum up, the results of the present analysis indicate that the texts that most closely matched the oral corpus in terms of FL were authentic materials drawn from news or television programmes. A future analysis comparing authentic materials and purpose-written textbook dialogues may reveal whether the two are indeed significantly different with respect to the number of lexical bundles. Many authors have pointed out the

importance of providing authentic input to language learners but the results of the current study reveal that overall, the FL in French textbooks is not representative of speech. Fortunately, it seems to be the case that publishers are making an effort to include more authentic materials in textbooks.

Correspondence should be addressed to Nathan Vandeweerd.
Email: Nathan.vandeweerd@uclouvain.be

Notes

[1]This does not have an effect on the overall number of definite articles in the lexical bundles extracted, only the gender of the definite article. After extraction, the gender of the definite article was corrected.

[2]The genres included texts, phrases, dialogues (and interviews), letters, mail, advertisements, poems, and recipes. Only the dialogues (and interviews) were used in the current study.

[3]All textbooks published by Difusión were excluded from the current study as no new textbook was published since 2008.

[4]Textbooks published by Didier were not included in the analysis because there were only three textbooks at the A2 level and one textbook at the B1 level.

References

Bare Bones Software Inc. (2016). TextWrangler. North Chemsford, MA. Retrieved from http://www.barebones.com/products/textwrangler/

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at . . .: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371-405. http://doi.org/10.1093/applin/25.3.371

Boers, F., & Lindstromberg, S. (2012). Experimental and intervention studies on formulaic sequences in a second language. *Annual Review of Applied Linguistics*, *32*, 83-110. http://doi.org/10.1017/S0267190512000050

Bolly, C. (2008). *Les unités phraséologiques : un phénomène linguistique complexe ? Séquences (semi-) figées construites avec les verbes "prendre" et "donner" en français écrit L1 et L2 : approche descriptive et acquisitionnelle* (Doctoral dissertation). Université catholique de Louvain, Louvain-la-Neuve, Belgium. Retrieved from http://hdl.handle.net/2078.1/19625

Branca-Rosoff, S., Fleury, F., Lefeuvre, F., & Pires, M. (2012). *Discours sur la ville. Corpus de Français Parlé Parisien des années 2000* (CFPP2000). Retrieved from http://cfpp2000.xn--univparis3-tt6e.fr/

Chambers, A. (2009). Les corpus oraux en français langue étrangère : authenticité et pédagogie. *Mélanges Crapel*, *31*, 1533. Retrieved from http://www.atilf.fr/IMG/pdf/melanges/01_Chambers.pdf

Chomsky, N. (1994). *Bare phrase structure*. Cambridge, MA: MIT Working Papers in Linguistics.

Clark, H. (1996). *Using language* ('Using' Linguistic Books). Cambridge, United Kingdom: Cambridge University Press.

CNRS. (2016). *COllections de COrpus Oraux Numériques*. Retrieved from http://cocoon.huma-num.fr/exist/crdo/

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics*, *32*, 45-61. http://doi.org/10.1017/S0267190512000074

Council of Europe. (2001). *The Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge, United Kingdom: Press Syndicate of the University of Cambridge.

Coveney, A. (2002). *Variability in spoken French: A sociolinguistic study of interrogation and negation*. Bristol, United Kingdom: Intellect Books.

Cruttenden, A. (1981). Item-learning and system-learning. *Journal of Psycholinguistic Research*, *10*(1), 79-88. http://doi.org/10.1007/BF01067363

Edmonds, A. (2014). Conventional expressions. *Studies in Second Language Acquisition*, *36*(01), 69-99. http://doi.org/10.1017/S0272263113000557

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics*, *32*, 17-44. http://doi.org/10.1017/S0267190512000025

Ellis, N. C., & Larsen-Freeman, D. (2009). Constructing a second language: Analyses and computational simulations of the emergence of linguistic constructions from usage. *Language Learning*, *59*, 90-125. http://doi.org/10.1111/j.1467-9922.2009.00537.x

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text - Interdisciplinary Journal for the Study of Discourse*, *20*(1), 29-62. http://doi.org/10.1515/text.1.2000.20.1.29

Forsberg, F. (2008). *Le langage préfabriqué : formes fonctions et fréquences en français parlé L2 et L1*. Bern, Switzerland: Peter Lang.

Forsberg, F. (2010). Using conventional sequences in L2 French. *IRAL - International Review of Applied Linguistics in Language Teaching*, *48*(1), 25-51. http://doi.org/10.1515/iral.2010.002

Forsberg, F., & Bartning, I. (2010). Can linguistic features discriminate between the communicative CEFR-levels?: A pilot study of written L2 French. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing* (pp. 133-157). European Second Language Association. Retrieved from http://eurosla.org/monographs/EM01/EM01tot.pdf

François, T. (2011). *Les apports du traitement automatique du langage à la lisibilité du français langue étrangère* (Doctoral dissertation). Université catholique de Louvain, Louvain-la-Neuve, Belgium). Retrieved from http://hdl.handle.net/2078.1/92162

François, T. (2014). An analysis of a French as a foreign language corpus for readability assessment. In E. Volodina, L. Borin, & I. Pilán (Eds.), *Proceedings of the 3rd workshop on NLP for computer-assisted language learning* (pp. 13-32). Linköping, Sweden: Linköping University Electronic Press. Retrieved from http://hdl.handle.net/2078.1/152560

Giradet, J., & Pécheur, J. (2010a). *Écho A2*. Paris, France: CLE International.

Giradet, J., & Pécheur, J. (2010b). *Écho B1 - Volume 2*. Paris, France: CLE International.

Google. (2016). Convert PDF and photo files to text. Retrieved from https://support.google.com/drive/answer/176692?hl=en

Gouverneur, C. (2008). The phraseological patterns of high-frequency verbs in advanced English for general purposes: A corpus-driven approach to EFL textbook analysis. In F. Meunier & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 223-243). Amsterdam, Netherlands: John Benjamins.

Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 27-50). Amsterdam, Netherlands: John Benjamins.

Gries, S. (2006). Exploring variability within and between corpora: Some methodological considerations. *Corpora*, *1*(2), 109-151. Retrieved from http://www.euppublishing.com/doi/pdf/10.3366/cor.2006.1.2.109

Ishihara, N., & Paller, D. L. (2016). Research-informed materials for teaching pragmatics, The case of agreement and disagreement in English. In B. Tomlinson (Ed.), *SLA research and materials development for language learning* [e-book] (pp. 87-102). New York, NY: Routledge.

Jiang, N., & Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *Modern Language Journal*, *91*(3), 433-445. http://doi.org/10.1111/j.1540-4781.2007.00589.x

Koprowski, M. (2005). Investigating the usefulness of lexical phrases in contemporary coursebooks. *ELT Journal*, *59*(4), 322-332. http://doi.org/10.1093/elt/cci061

Kuiper, K. (2004). Formulaic performance in conventionalised varieties of speech. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 37-54). Amsterdam, Netherlands: John Benjamins.

Laboratoire ICAR. (2014). CLAP: Corpus de langues parlées en interaction. Retrieved from http://clapi.ish-lyon.cnrs.fr/

Lundell, F. F., & Lindqvist, C. (2012). Vocabulary aspects of advanced L2 French: Do lexical formulaic sequences and lexical richness develop at the same rate? *Language, Interaction and Acquisition/Langage, Interaction et Acquisition*, *3*(1), 73-92. http://doi.org/10.1075/lia.3.1.05for

MacWhinney, B. (2012). The CHILDES Project Tools for Analyzing Talk-Electronic Edition Part 1: The CHAT Transcription Format. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.359.4451

Myles, F., Hooper, J., & Mitchell, R. (1998). Rote or rule? Exploring the role of formulaic language in classroom foreign language learning. *Language Learning*, *48*(3), 323-363. http://doi.org/10.1111/0023-8333.00045

Myles, F., Mitchell, R., & Hooper, J. (1999). Interrogative chunks in French L2: A basis for creative construction. *Studies in Second Language Acquisition*, *21*, 49-80. http://doi.org/10.1017/S0272263199001023

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 191-226). New York, NY: Routledge.

Raupach, M. (1984). Formulae in second language speech production. In H. W. Dechert, D. Mohle, & M. Raupach (Eds.), *Second language productions* (pp. 114-137). Tubingen, Germanu: Gunter Nar.

Suethanapornkul, S. (2009). A story in four words: An analysis of lexical bundles in learners' writing placement test in ELIPT corpus. Retrieved from http://scholarspace.manoa.hawaii.edu/handle/10125/20172

Tomlinson, B. (2016). Achieving a match between SLA theory and materials development. In B. Tomlinson (Ed.), *SLA research and materials development for language learning* [e-book] (pp. 3-22). New York, NY: Routledge.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, *17*(1), 84-119. http://doi.org/10.1093/applin/17.1.84

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning*, *61*(2), 569-613. http://doi.org/10.1111/j.1467-9922.2010.00622.x

Tsai, K. (2015). Profiling the collocation use in ELT textbooks and learner writing. *Language Teaching Research*, *19*(6), 723-740. http://doi.org/10.1177/1362168814559801

Underwood, G., Schmitt, N., & Galpin, A. (2004). The eyes have it: An eye-movement study into the processing of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 153-172). Amsterdam, Netherlands: John Benjamins.

Webb, S., Newton, J., & Chang, A. (2013). Incidental learning of collocation. *Language Learning*, *63*(1), 91-120. http://doi.org/10.1111/j.1467-9922.2012.00729.x

Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics*, *21*(4), 463-489. http://doi.org/10.1093/applin/21.4.463

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge, United Kingdom: Cambridge University Press.

Wray, A., & Perkins, M. R. (2000). The functions of formulaic language: An integrated model. *Language & Communication*, *20*(1), 1-28. http://doi.org/10.1016/S0271-5309(99)00015-4